Chapter 1
Introduction to Statistics and Data Visualisation

Εἰκὸς γὰρ γίνεσθαι πολλὰ καὶ παρὰ τὸ εἰκός.
It is likely that unlikely things should happen.
Aristotle, Poetics, 1456a, 24

Although it is a common perception that statistics seeks to quantify and categorise uncertainty and unlikely events, it is actually a much broader and more general field. In fact, statistics is the science of collecting, analysing, interpreting, and displaying data in an objective manner. Built on a strong foundation in probability, the application of statistics has expanded to consider such topics as curve fitting, game theory, and forecasting. Its results are applied in many different fields, including biology, market research, polling, economics, cryptography, chemistry, and process engineering.

Basic statistical methods have been traced back to the earliest times in such forms as the collection of data regarding a farmer's livestock; the amount, quality, and type of grain in the city granaries; or the phases of the moon by early astronomers. With these simple data sets, graphs could be created, summary values could be computed, and patterns could be detected and used. Greek philosophers, such as Aristotle (384–322 B.C), pontificated on the meaning of probability and its different realisations. Meanwhile, ancient astronomers, such as Ptolemy (*c.* A.D. 90–168) and Al-Biruni (973–1048), were developing methods to deal with the randomness and inherent errors in their astronomical measurements. By the start of the late Middle Ages around 1,300, rudimentary probability was being developed and applied to break codes. With the start of the seventeenth century and spurned by a general interest in games of chance, the foundations of statistics probability were developed by Abraham de Moivre (1667–1754), Blaise Pascal (1623–1662), and Jacob Bernoulli (1655–1705). These scientists sought to resolve and determine optimal strategies for such games of chance. The nascent nation states also took a strong interest in the collection and interpretation of economic and demographic information. In fact, the word *statistics*, first used by the German philosopher Gottfried Achenwall (1719–1772) in 1749, is derived from the Neolatinate term *statisticum collegium*, meaning *council of the state*, referring to the fact that even then the primary use of the collected information was to provide insight (*council*) about the nation state (Varberg 1963). In the early nineteenth century, work by

Y.A.W. Shardt, *Statistics for Chemical and Process Engineers*,
DOI 10.1007/978-3-319-21509-9_1

amongst others Johann Carl Friedrich Gauss (1777–1855), Pierre-Simon Laplace (1749–1827), and Thomas Bayes (1701–1761) led to the development of new theoretical and practical ideas. Theoretically, the grounding of statistics in probability theory, especially the development of the Gaussian distribution, allowed for many practical applications, including curve fitting and linear regression. Subsequent work, by such researchers as Andrei Kolmogorov (1903–1987) and Andrei Markov (1856–1922), solidified the theoretical underpinning and developed new ways of understanding randomness and methods for quantifying its behaviour. From these foundations, Karl Pearson (1857–1936) and Ronald Fisher (1890–1962) developed hypothesis testing, the χ^2-distribution, principal component analysis, design of experiments, analysis of variance, and method of maximum likelihood, which continue to be used today. Subsequently, these ideas were used by George Box (1919–2013), Gwilym Jenkins (1932–1982), and Lenart Ljung (1946−) to develop stochastic modelling and advanced probabilistic models with applications in economics, biology, and process control. With the advent of computers, many of the previously developed methods can now be realised efficiently and quickly to analyse enormous amounts of data. Furthermore, the increasing availability of computers has led to the use of new methods, such as Monte Carlo simulations and bootstrapping.

Even though statistics still remains solidly applied to the study of economics and demographics, it has broadened its scope to cover almost every human endeavour. Some of the earliest modern applications were to the design and analysis of agricultural experiments to show which fertilisers and watering methods were better despite uncontrollable environmental differences, for example, amount of sunlight received or local soil conditions. Later these methods were extended to analyse various genetic experiments. Currently, with the use of powerful computers, it is possible to process and unearth unexpected statistical relationships in a data set given many thousands of variables. For example, advertisers can now accurately predict changes in consumer behaviour based on their purchases over a period of time.

Another area where statistics is used greatly is the chemical process industry, which seeks to understand and interpret large amounts of industrial data obtained from a given (often, chemical) process in order to achieve a safer, more environmentally friendly, and more profitable plant. The process industry uses a wide range of statistics, ranging from simple descriptive methods through to linear regression and on to complex topics such as system identification and data mining. In order to appreciate the more advanced methods, there is a need to thoroughly understand the fundamentals of statistics. Therefore, this chapter will start the exploration with some fundamental results in statistical analysis of data sets coupled with a thorough analysis of the different methods for visualising or displaying data. Subsequent chapters will provide a more theoretical approach and cover more complex methods that will always come back to use the methods presented here. Finally, as a side note, it should be noted that the focus of this book is on presenting methods that can be used with modern computers. For these reasons, heavy emphasis will be made on matrices and generalised approaches to solving the problems. However, except for

the last two chapters dedicated to MATLAB® and Excel®, little to no emphasis will be placed on any specific software as a computational tool; instead the theoretical and implementation aspects will be examined in depth.

1.1 Basic Descriptive Statistics

The most basic step in statistical analysis of a data set is to describe it descriptively, that is, to compute properties associated with the data set and to display the data set in an informative manner. A data set consists of a finite number of *samples* or data points. In this book, a data set will be denoted using either set notation, that is, $\{x_1, x_2, \ldots, x_n\}$ or vector notation, that is, as $\vec{x} = \langle x_1, x_2, \ldots, x_n \rangle$. Set notation is useful for describing and listing the elements of a data set, while vector notation is useful for mathematical manipulation. The size of the data set is equal to n. The most common descriptive statistics include measures of *central tendency* and *dispersion*.

1.1.1 Measures of Central Tendency

Measures of central tendency provide some information about the most common value in the data set. The basic measures of central tendency include the *mean*, *mode*, and *median*. Since the most common such measure is the *mean*, which is often colloquially called the average, all of these measures are often referred to as *averages*. A summary of the basic properties of these measures is provided in Table 1.1.

The *mean* is a measure of the central value of the set of numbers. It is often denoted as an overbar ($\overline{\circ}$) over a variable, for example, the mean of $\vec{x}$ would be written as $\overline{x}$. The most common **mean** is simply the sum of all the values divided by the total number of data points, n, that is,

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

Alternatively, a weighted mean can be computed, where for each value a weight w is assigned, that is,

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{1.2}$$

Table 1.1 Summary of the main properties of the measures of central tendency

Measure	Formula	Advantages	Disadvantages
Mean	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	Easy to compute and interpret	Can easily be influenced by extreme values
Mode	Most common entries in the data set	Easy to interpret	Many not accurately represent the data set
Median	Middle entry of the ordered data set	Robust and easy to interpret	Not necessarily easy to compute

The weighted mean can be used when the accuracy of some of the values is suspected to be less than that of others. Although the mean is a commonly used measure of central tendency and hence widely reported when describing data, it is not necessarily a robust measure, that is, the mean can be heavily skewed by one or two numbers that are significantly different from the others. For example, if we have the data set of three numbers, {2, 3, 4}, whose mean is $\overline{x} = 3$, and replace the 4 by 10, the mean becomes $\overline{x} = 5$, which is larger than two of the other numbers.

The *mode* represents the most common entry in a given data set. Multiple entries can be tied for the mode, in which case, the data set is said to be *multimodal*.[1] For the following set of numbers, {2, 4, 5, 5, 5, 6, 7, 10, 10, 10, 11}, there are two modes: 5 and 10, as both occur exactly 3 times. Although, in general, the mode is less sensitive to minor changes in the data set, it is still relatively easy to skew the results by adding too many identical values to create a new modal value. Furthermore, the most common entry need not be in any way descriptive of the overall properties of the data set. This can especially be the case if one of the extreme values occurs slightly more often than the other numbers and hence becomes the modal value.

The *median* represents the middle value of an ordered data set. If the number of data points is odd, then the median will represent the middle value. On the other hand, if the number of data points is even, then the median will be the mean value of the two middle values. Although it can happen that the median value is equal to a value in the data set, this is not necessarily always true. For the set given as {2, 4, 5, 10, 14, 14, 16, 17}, the median value would be 12 (= ½(10 + 14)). The main advantage of the median value is that it represents the middle value of a given set and is robust to single extreme values.

1.1.2 Measures of Dispersion

Measures of dispersion seek to provide some information about how the values in a given data set are distributed, that is, are all the values clustered about one number

[1] If the specific number of tied entries is known, then the data set can be referred to by that number, for example, *bimodal* for a data set with 2 modes or *trimodal* for three modes.

Table 1.2 Summary of the main properties of the measures of dispersion

Measure	Formula	Advantages	Disadvantages	Comment
Range	Max – min or [min, max]	Easy to compute	Can easily be influenced by extreme values	
Standard deviation	$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$	Commonly used, can be easy to interpret	Can easily be influenced by extreme values	Squaring it gives the variance
Median absolute difference	$\hat{\sigma}_{MAD} = \text{median}(\lvert x_i - \overline{x}_{median} \rvert)$	Robust estimate		Can be converted to an estimate of the standard deviation
Skew	$\hat{\gamma} = \frac{n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1.5}}$	Measures the spread of the extreme values		Rarely used in practice

or are they spread out across a large range of numbers. The basic measures of dispersion include *range*, *standard deviation* or *variance*, *skew*, and *median absolute deviation* (*MAD*). A summary of the basic properties of these measures is provided in Table 1.2.

The *range* of a data set is simply defined as the difference between the largest and smallest values within the data set. It is also possible to report the range as the two numbers representing the extreme data set values. It provides a simple, but not very meaningful, interpretation of the spread of the values. The larger the range, the more spread out the values are. Clearly, the range is affected adversely by large extreme values, since they would be directly used in its computation.

The *standard deviation*, σ, and *variance*, σ^2, are two related measures of the spread of the data set. The variance is always equal to the second power of the standard deviation. The larger the standard deviation, the more spread out the data set is. The variance can be computed as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{1.3}$$

The standard deviation can then be computed by taking the square root of the value obtained using Eq. (1.3). In statistics, the circumflex (ˆ) over a value shows that it is estimated or computed from a data set, rather than some theoretical value, for example, in Eq. (1.3), $\hat{\sigma}^2$ is the estimated value of the true variance, σ^2, given the data set. Even if the variance for the data set were the same, taking different data points will lead to some variation in the computed value. It can be noted that the

variance is sensitive to extreme values. Occasionally, the variance can be denoted as the function *var*, for example, var(x) is the variance of x.

A method to avoid the sensitivity of the standard deviation to extreme values is to compute the *median absolute deviation (MAD)*, denoted by σ_{MAD}, which replaces the mean by the robust median. It can be computed as follows:

$$\hat{\sigma}_{\text{MAD}} = \text{median}(|x_i - \overline{x}_{\text{median}}|) \tag{1.4}$$

where *median* is the function that determines the median value given a data set and $\overline{x}_{\text{median}}$ is the median value for the data set. It is possible to convert $\hat{\sigma}_{\text{MAD}}$ to a robust estimate of the standard deviation. However, it requires knowing the underlying distribution in order to compute the conversion factor. For a normal distribution, the robust estimate of the standard deviation can be written as

$$\hat{\sigma} = 1.4826\hat{\sigma}_{\text{MAD}} \tag{1.5}$$

The *skew*, denoted by γ, measures the amount of asymmetry in the distribution. Skewness is determined by examining the relationship in the clustering of extreme values, that is, the tails. If more of the data set is clustered towards the smaller extreme values, then it is said that the system has *positive* or *right skewness*. On the other hand, if the data set is clustered towards the larger extreme values, then it is said that the system has *negative* or *left skewness*. The skew of a data set can be computed as

$$\hat{\gamma} = \frac{n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1.5}} \tag{1.6}$$

Graphically, the skewness can be seen from a histogram, which plots the frequency of a value against the value. Examples of left and right skewness are shown in Fig. 1.1.

1.1.3 Other Statistical Measures

In addition to the measures of central tendency and dispersion, there exist other ways of quantifying a particular data set. This section will briefly review the two most common such methods: *quantiles* and *outliers*.

1.1.3.1 Quantiles

A *quantile* is a way of dividing the data set into segments based on the ordered rank of the data set. Common quantiles are the median (2 segments with the split at

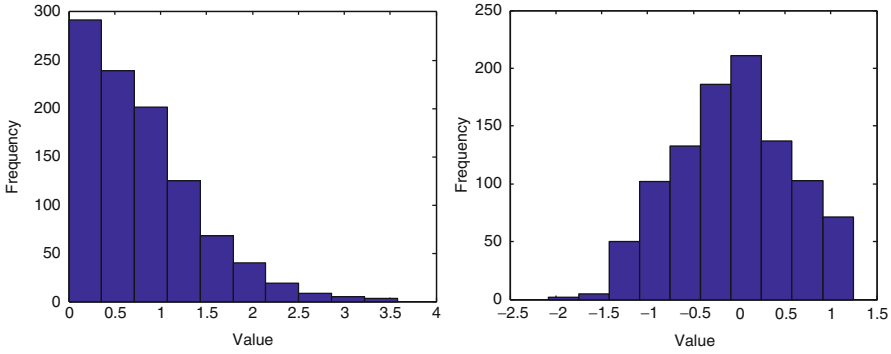


Fig. 1.1 (*Left*) Right-skewed and (*right*) left-skewed data set

50%), *quartiles* (4 segments at 25, 50, and 75%), *quintiles* (5 segments at 20, 40, 60, and 80%), and *percentiles* (100 segments). In order to obtain a meaningful division, there should be at least as many different data points as there are segments.

Partitioning a data set into quantiles can be accomplished using the following steps:

1. *Order* the data set from smallest to largest.
2. *Obtain* an estimate of the boundaries for each of the segments using the following formula (Hyndman and Fan 1996)

$$\begin{aligned} h &= (n-1)\,p + 1 \\ Q_p &= x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\left(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor}\right) \end{aligned} \tag{1.7}$$

where n is the number of data points, $p = k/q$, k, and q are defined as the kth q-tile, x_i is the ith data point of the ordered data set, and $\lfloor \cdot \rfloor$ is the floor function, that is, round down any number to its nearest integer. When $p = 1$, then $Q_p = x_n$.

Different formulae for computing the sample quantile can be obtained by changing the equation for h. Two other common formulations are:

1. *Exclusive Formulation*: $h = (n+1)p + 1$ with both p and Q_p computed the same way as before.
2. *Linear Interpolation Formulation*: $h = np + ½$ with both p and Q_p computed the same way as before.

The differences in the estimated values are in most cases quite minimal. A comparison of the above methods is given in Table 1.6 in the context of an example (see Sect. 1.3.3, p. 28).

It can be noted that in all versions of Excel®, the method defined by Eq. (1.7) is available (as either the function `quartile` or `quartile.inc`). Newer versions of Excel® (2010 or newer) also support option 1 (as `quartile.exc`). All versions of MATLAB® implement option 2.

1.1.3.2 Outliers

Outliers are data points that seem to be quite different from surrounding values and expected behaviour. Outliers can be caused by many different factors, including data entry or data collection errors or caused by randomness inherent in the system. Whenever a point is suspected to be an outlier, it is always useful to check that it has been correctly recorded and collected. Determining whether a point is an outlier is ultimately subjective and relies on intuition. Common rules for determining outliers include (Lin et al. 2007):

1. *Visual Tests*: visual inspection to determine which values are located far from the bulk of the data, for example, in the set $\{1, 2, 1, 2, 3, 0, 2, -10\}$, -10 could be considered to be an outlier. Displaying the data using graphs can be a very useful approach. Graphs, such as the box-and-whisker plot, line charts, and scatter plots, can be useful for determining outliers.
2. *3σ Edit Rule*: data points whose Z-score are large (>3), where the Z-score is given as

$$Z_i = \frac{x_i - \overline{x}}{\sigma} \tag{1.8}$$

   x_i is the data point of interest, Z_i is the corresponding Z-score, $\overline{x}$ is the mean value of the data set, and σ is the standard deviation of the data set. This approach only works if it can be assumed that the data set comes from a normal distribution and is not very robust.
3. *Hampel identifier* (Davies and Gather 1993): the Hampel identifier assumes that points which lie outside the band $x_{\text{median}} \pm 3\sigma_{\text{mad}}$, where σ_{mad} is defined as

$$\sigma_{\text{mad}} = 1.4826\text{median}(|x_i - x_{\text{median}}|) \tag{1.9}$$

   and *median* is the function that determines the median value of the given data set. This equation represents the median absolute difference and is a robust manner of estimating the spread of the data. The constant is selected such that σ_{mad} is equal to σ for a normal distribution. In fact, for a normal distribution, the Hampel identifier and the 3σ edit rules will produce the same results.

1.2 Data Visualisation

Data visualisation is the science and art of displaying information in a visual manner that not only displays the relevant information accurately but is also visually appealing. There exist many different methods for visualising a given data set, including graphs and tables. Each method has its advantages and disadvantages when it comes to displaying the data. In general, the following principles can be followed to determine which method is best to display the data:

1. *Density of Information*: how much information is to be presented? Are there only a few points that need to be summarised, or are there multiple points that need to be shown?
2. *Comparison*: what is the point of showing the values? What types of relationships between the data are to be highlighted?
3. *Efficiency*: which method shows the desired relationships the best? How well is the information displayed? Are the desired relationships visible clearly?
4. *Display Scheme*: what kind of display scheme will be required? Will you need to use different colours? If so, how many? Will you need to use multiple different symbols? If so, which ones? Can they all be distinguished easily in the figure? What if the figure is printed in black and white? What type of scale will be used: normal or logarithmic?

Irrespective of the method selected, it is important that the following information, as appropriate, be included:

1. *Titles/Captions*: each figure or group of figures should have a clear title or caption that briefly explains the information in the figure.
2. *Labels*: appropriate labels should be included. This should include, as appropriate, the full name of what is being shown, abbreviations, and units. All axes and legend headings should be considered. For axes, an acceptable and very useful approach would be to use the following label "full name, abbreviation (units)", for example, "temperature, T (°C)". A legend should be provided if multiple types of information are plotted on the same graph.
3. *Display*: are the different symbols used clearly distinguishable? Consider the fact that many figures will end up in black-and-white publications. This implies that relying solely on colour to distinguish different aspects on a figure can be difficult. Furthermore, data points should not be connected by lines unless there is a reason for connecting the points. This implies that experimental data in many cases should be entered as single points, while theoretical values should be connected with a single continuous line.

A good discussion of the art of data visualisation, as well as some ideas on how to implement it, can be found in the books by Edward Tufte (Tufte 2001; Tufte 1997).

1.2.1 Bar Charts and Histograms

A *bar chart* is a graph that contains vertical or horizontal bars whose length is proportional to the value. Bar charts compare by their nature discrete information. One axis will contain the category or discrete item, while the other axis will contain the value axis. Typical bar charts are shown in Fig. 1.2. Although 3-D bar charts are possible, they do not provide any advantage for displaying the information accurately or efficiently.

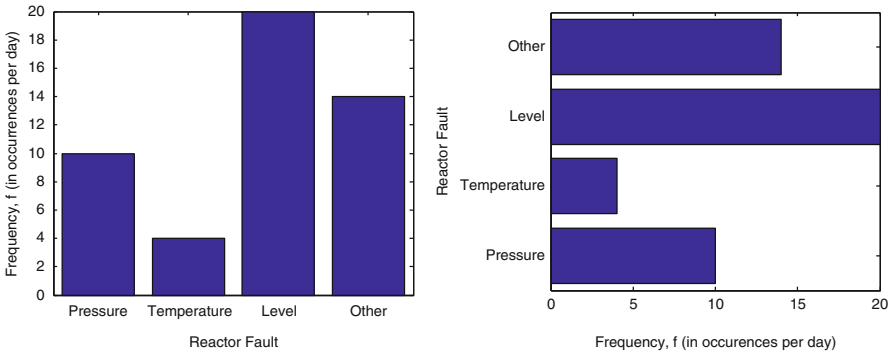


Fig. 1.2 (*Left*) Vertical bar chart and (*right*) horizontal bar chart

A *histogram*, similar to a bar chart, shows the frequency of a given range of values that occur in the data set. Thus, a histogram records continuous data but presents it in a similar manner. A histogram is constructed by first creating bins or ranges of numbers. Next, the number of times a value from the data set falls within each of the ranges is determined and noted. Once this has been completed, a vertical bar chart is plotted using the bins as the category and the occurrences as the value. It should be noted that the bins are normally assigned so that they are of equal size (except for the two end points) and are continuous, that is, two adjacent bins share the same end point. A 4-bin example could be $x < 3$, $3 \leq x < 5$, $5 \leq x < 7$, and $x \geq 7$. A typical histogram is shown in Fig. 1.3. Not all software provides methods for directly creating a histogram. In some cases, it is necessary to manually bin the data and then create the corresponding bar graph.

1.2.2 Pie Charts

A *pie chart* is a circle whose arc length has been divided up into different proportions. It is named after how a pie is cut. Pie charts can be used to display the relationships of parts to a whole, for example, components of a budget. However, too many different items in a pie chart can lead to difficulties with representing the items effectively, as the number of available colours and amount of space can be limited. Also, a pie chart tends to require more space than would ideally be needed to display the information. A typical pie chart is shown in Fig. 1.4.

1.2.3 Line Charts

A *line chart* is a graph that contains individual data points connected by a line. Very often, the horizontal, or x-axis, will represent time and the vertical, or

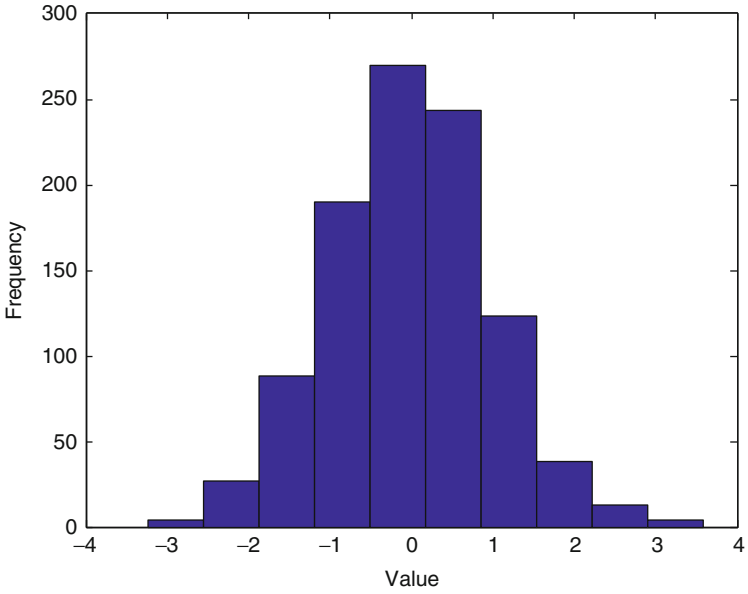
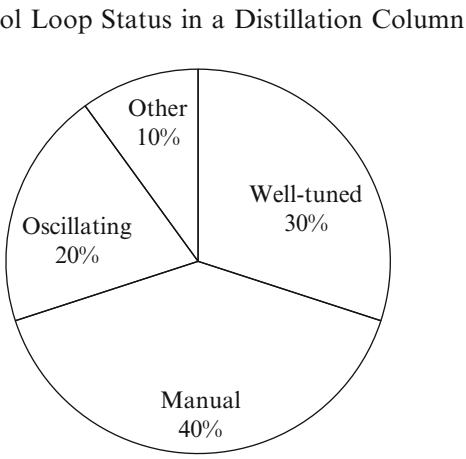


Fig. 1.3 Typical histogram

Fig. 1.4 Typical pie chart



y-axis, will represent the value of some variable over time. For this reason, a line chart is often called a *time series plot*. A line chart is very effective in showing how a variable(s) changes over time. However, too many competing lines can make the figure difficult to read and understand. A typical line chart is shown in Fig. 1.5.

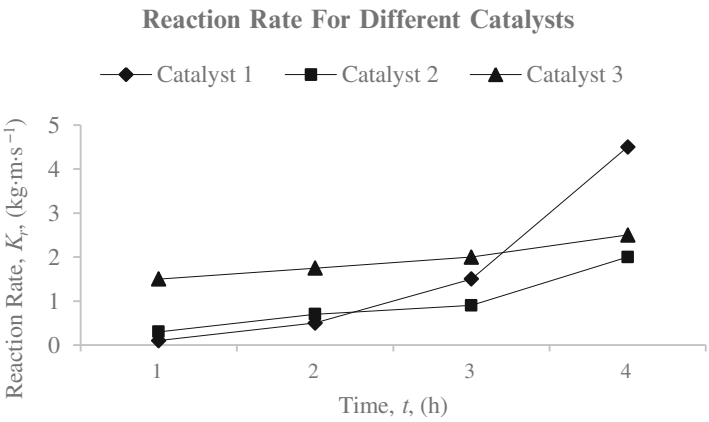


Fig. 1.5 Typical line chart

1.2.4 Box-and-Whisker Plots

A *box-and-whisker plot*, or more simply a *boxplot*, is a complex graph that is based on quartiles to conveniently display multiple different properties of the data set. It can conveniently be used to compare different data sets. A box-and-whisker plot consists of two parts: the box and the whiskers. The box is formed by the 25th (Q1) and 75th (Q3) percentile boundaries with the middle line invariably being the median (Q2). The whisker limits are defined using any of the following rules:

1. Maximum and minimum of the data set.
2. Lowest data point located within 1.5 of the interquartile range from the lower quartile and the largest data point located within 1.5 of the interquartile range above the upper quartile. The interquartile range is defined as the difference between Q3 and Q1. Such a plot is often called a *Tukey boxplot*.
3. The 9th and 91st percentiles.
4. The 2nd and 98th percentiles.

In all cases, data points lying outside the whisker limits are conventionally denoted by crosses or dots, often in another colour. Such points can be labelled as *outliers*. Of the available definitions, the most commonly encountered box-and-whisker plots use whisker bounds defined by the first two rules. Typical box-and-whisker plots are shown in Fig. 1.6. These box-and-whisker plots were created using the interquartile range for the data points.

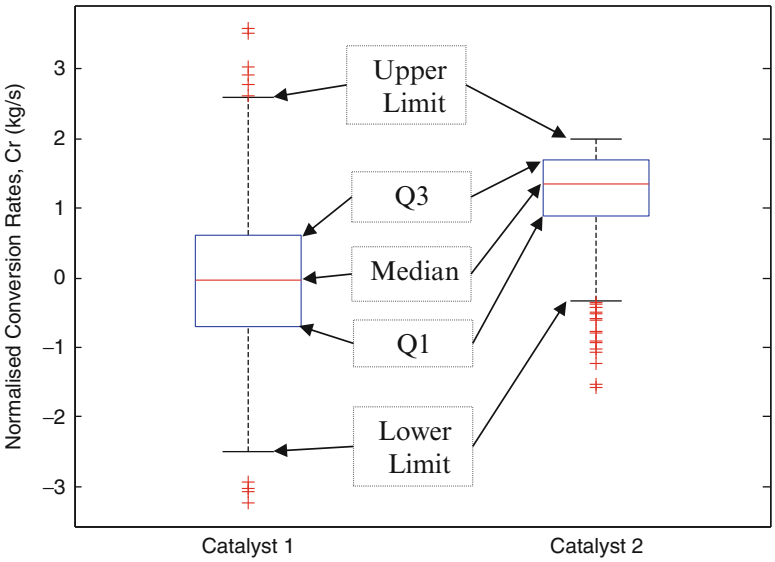


Fig. 1.6 Typical box-and-whisker plots

1.2.5 Scatter Plots

A *scatter plot* shows the values obtained using some mark. These marks are not connected and hence it looks like all the values are scattered around. A scatter plot is useful when it is desired to show the relationship between two variables, but the values vary quite a bit between each sample. Often, the true or predicted values can be superimposed using a line. The selection of the appropriate mark can be important, especially when there are many data points to show. Ideally, each data point should be clearly visible. In some cases, it may be useful to show data from multiple experiments or runs together on a single plot. Again, the various marks need not only to be individually distinguishable, but also they need to be distinguishable from each other. A typical scatter plot is shown in Fig. 1.7.

1.2.6 Probability Plots

A *probability plot* is a graph that compares the data set against some expected statistical distribution by comparing the actual quantiles against the theoretical quantiles. Such probability plots are also often called $Q-Q$ or $P-P$ plots. The most common statistical distribution for comparison is the normal distribution. The exact values plotted on each of the axes depend on the desired graph and software used. In general, the theoretical values are plotted on the x-axis, while the actual values are plotted on the y-axis. Occasionally, the actual values are modified in

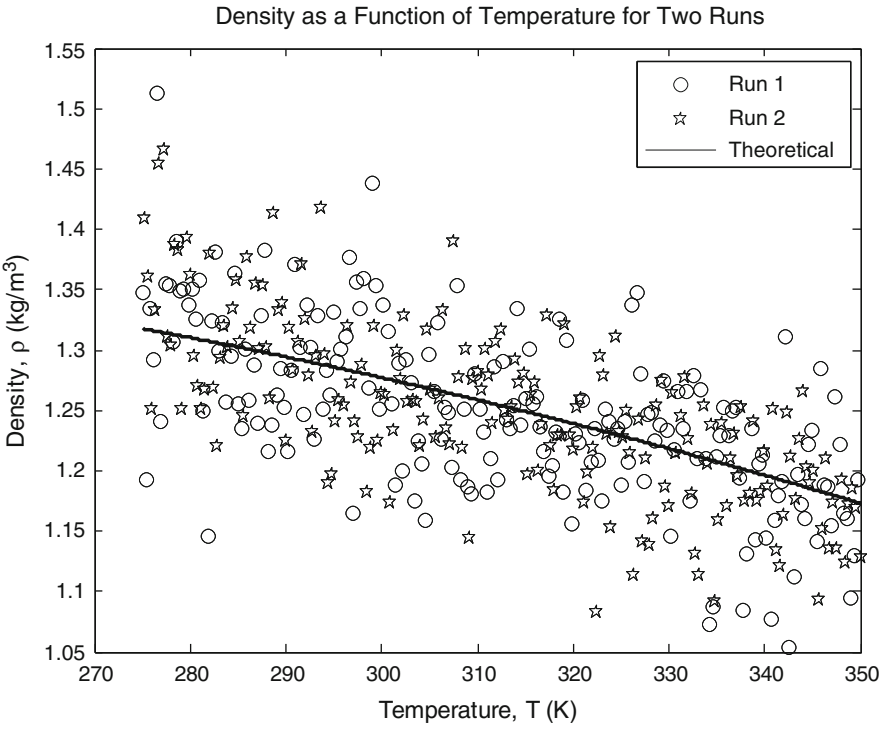


Fig. 1.7 Typical scatter plot

order to emphasise certain properties. A generalised probability plot can be constructed using the following steps:

1. For each data point, compute its rank, denoted by R_i.
2. Compute an approximation of the quantile position using the following formula:

$$U_{Ri} = \begin{cases} 1 - 0.5^{\frac{1}{n}} & i = 1 \\ \dfrac{i - 0.3175}{n + 0.365} & i = 2, 3, \ldots, n-1 \\ 0.5^{\frac{1}{n}} & i = n \end{cases} . \tag{1.10}$$

It can be noted that any of many different formulae can be used here. The simplest formula is given as

$$U_{Ri} = \frac{i - 0.5}{n} . \tag{1.11}$$

The final results will be very similar, irrespective of the actual formula used.

3. Compute the N-score for each rank, N_{Ri}, using the following formula:

$$N_{Ri} = \text{icdf}(U_{Ri}) \tag{1.12}$$

where *icdf* is the inverse of the cumulative distribution function of the desired distribution (further information about distribution functions can be found in Sect. 2.3).

4. If desired, transform the actual data. Two common transformations are:

   (a) *Z-Score*: if the theoretical distribution is assumed to be normal, then it can be convenient to transform the data into the corresponding Z-score. This will minimise the need to know the true mean and standard deviation of the data set. The formula for the Z-score is

$$Z_i = \frac{x_i - \overline{x}}{\hat{\sigma}}. \tag{1.13}$$

   (b) *Quantiles*: another option is to plot the quantiles corresponding to the data set on the y-axes rather than the actual values. Any of the formulae for computing the quantile can be used. The most common one in this case is Eq. (1.11). This will give a cumulative distribution feel to the data set. Some software, such as MATLAB®, uses this approach to produce its probability plots.

5. Plot N_{Ri} on the x-axis and x_i on the y-axis to construct the normal probability plot.

The interpretation of this probability plot is based on the following theoretical observations:

1. The data should lie on a straight line, which, in the ideal case, is $y = x$.
2. If the straight line given by the data is shifted vertically by a fixed amount, then this represents the difference in the mean between the assumed distribution and the actual data distribution.
3. If the straight line given by the data has a different slope ($\neq 1$), then the standard deviation of the data set is different from the assumed distribution's standard deviation.

This is shown graphically in Fig. 1.8, for the case of a normal distribution with different means and variances compared against a normal distribution with a mean of zero and a variance of 1. It can be seen that the straight line's slope and y-intercept match well the theoretical values. Therefore, based on these observations, it can be useful to include a straight line (line of best fit) to give an estimate of the true mean and standard deviation.

From these theoretical observations, this means that the points in the probability plot should all lie along a straight line. The exact slope and y-intercept are not all that important. Deviations from a straight line are indications that the data may not come from the proposed theoretical distribution. The most common deviations are:

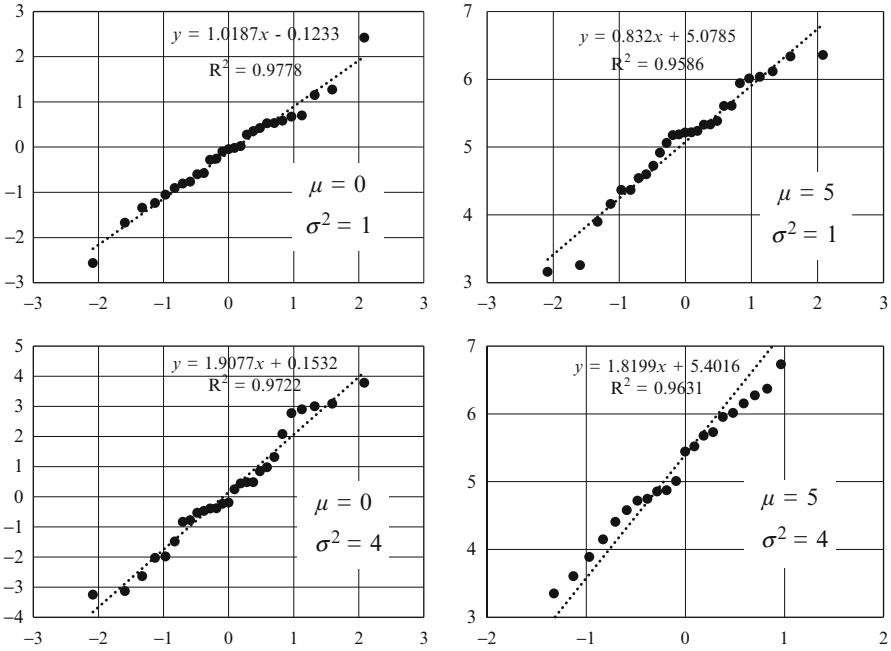


Fig. 1.8 Probability plots and the effect of the location parameters (μ and σ^2)

1. *Outliers* or extreme values at the end points.
2. *Tails at the end points*, or curvature, that is, one tail is below the straight line and the other is above the straight line. This implies that the true distribution of the data set has a different distribution than the target distribution. Practically, if the left tail is below and the right tail is above, then the distribution in the tails is larger than in the target distribution. On the other hand, if the left tail is above and the right tail is below the straight line, then the data distribution in the tails is smaller than in the target distribution.
3. *Convex or concave curvature* in the centre, that is, the given data set is not symmetric compared with the target distribution.
4. Plateaus, gaps, or horizontal data, that is, the data seems to fall only within certain values. This is most likely to be the result of rounding errors introduced during measurement, storage, or data manipulation.

Figure 1.9 shows examples of how these kinds of problems can appear on a probability plot. Figure 1.9a shows a normal probability distribution with mean 0 and variance 1 with 2 outliers (circled). Notice how the outliers can cause some of the adjacent points to also be skewed from the ideal location. Figure 1.9b shows the case where the tails of the distribution do not match. In this case, a 2-degree-of-freedom Student's *t*-distribution was compared against the normal distribution. The *t*-distribution has larger tails than the normal distribution. This can clearly be seen by the deviations on both sides from the central line. Figure 1.9c shows the case

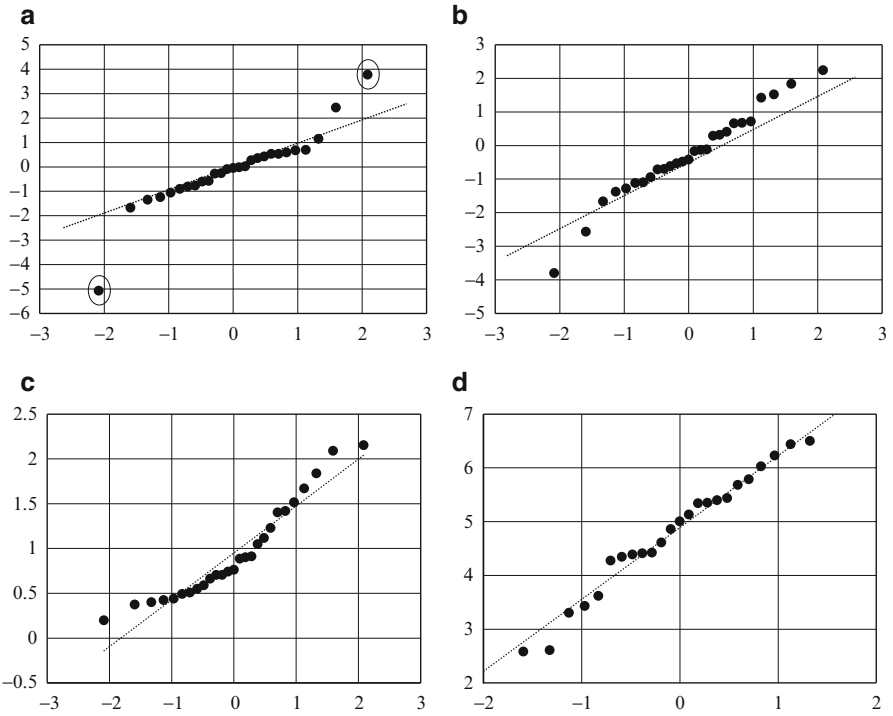


Fig. 1.9 Issues with probability plots. (**a**) Outliers. (**b**) Tails. (**c**) Concave behaviour. (**d**) Rounded to 3 decimal places

where there is convex curvature in the centre. In this case, the asymmetric F-distribution was compared with the normal distribution. In such a case, drawing the desired straight line can be quite difficult since there can potentially be two or more "best" regions. Figure 1.9 shows the case where there are horizontal plateaus combined with gaps. In this case, the normal distribution with mean of 5 and variance of 4 was rounded down to 3 decimal places. This clearly shows the gaps and plateaus that rounding can induce in the results. Furthermore, it should be noted that drawing the straight line for comparison can be difficult when the data set does not match the underlying distribution. Finally, when dealing with small samples (say less than about 30 points), then less ideal behaviour in the extreme regions (tails) can be tolerated. The extent and amount of tolerated deviations will depend on where the normal probability plot is being used. Figure 1.10 shows the normal probability plot for nine different realisations of eight data points drawn from the standard normal distribution. It can be seen that all samples show varying amounts of curvature and tails. Detailed comparisons of the effect of data size on normal probability plots can be found in (Daniel and Wood 1980).

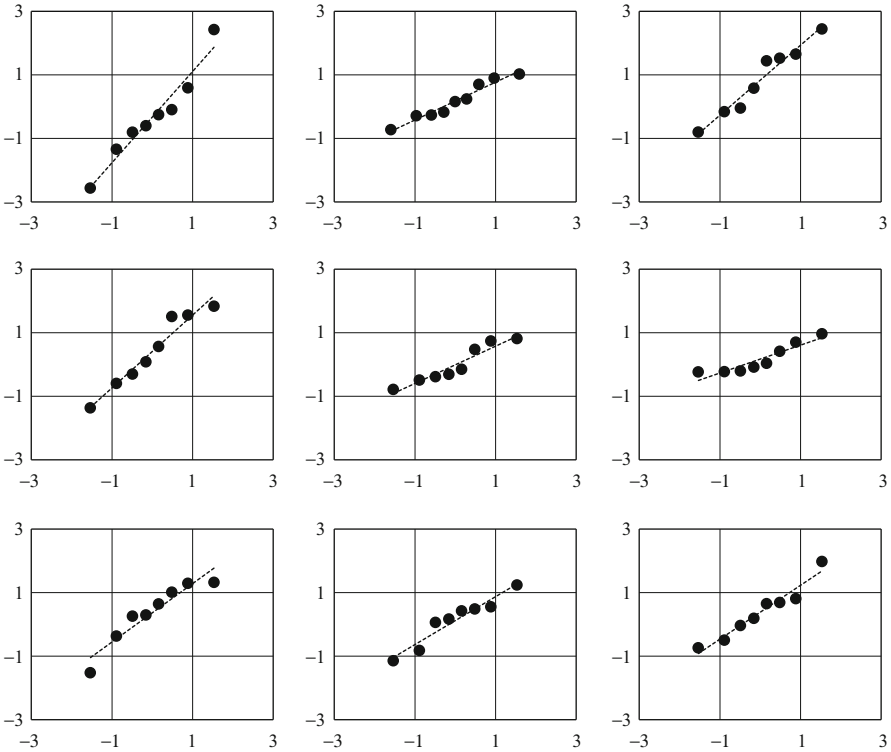


Fig. 1.10 Nine probability plots of eight samples drawn from a standard normal distribution

1.2.7 Tables

A *table* is a textual way of displaying information that consists of rows and columns. A table is useful to present a small amount of data whose exact values are important. It can be used to give information about summary statistics, such as the mean, mode, and standard deviation. Every table should have headers for its columns and rows. This can be formatted similarly to graph axes, by including the name of the variable, its symbol, and its units. A well designed table will contain all the relevant information within it and be self-explanatory. Numbers should be properly formatted and not taken straight from the software used. There is no need to display more than about 3 or 4 digits (unless special circumstances warrant) with spacing between groups of 3 digits (on both sides of the decimal place). Scientific notation should be used as appropriate, for example, the number obtained from a calculator as 1.25896321532e3 could be written as either 1.259×10^3 (using scientific notation) or 1,259. A typical table is shown in Table 1.3.

Table 1.3 Typical table formatting

Treatment	Mean thickness δ (μm)	Variance σ^2 (μm^2)	Range [lower, upper] (μm)
A	1.25	0.25	[0.25, 5.00]
B	1.50	0.10	[0.50, 2.25]
C	2.25	0.50	[0.50, 10.0]

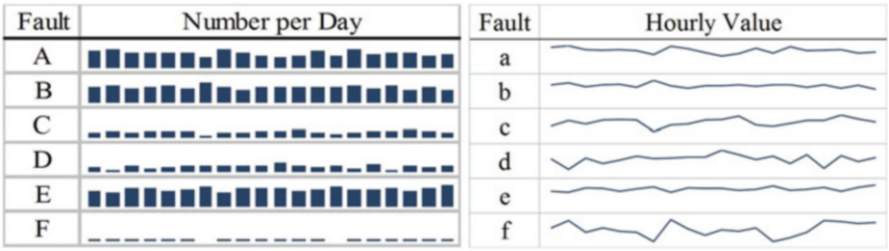


Fig. 1.11 (*Left*) Spark bar graph showing the number of times a given fault occurs over the course of many days and (*right*) sparkline showing the hourly process value for six different variables from a single unit over the course of a day

1.2.8 Sparkplots

Sparkplots or *profile plot* are various ways of summarising information so that only the trends and comparison between different data sets are compactly shown. Sparkplots often do not have explicit axes or category markings. Sparkplots can be either line graphs (known as sparklines) or bar graphs (known as spark bar graphs). It is common to use sparklines to show the behaviour of multiple process variables in order to understand which variables could be influencing others. Spark bar graphs are often used as histograms to show the distribution of variables and at the same time show the individual values. Typical examples are shown in Fig. 1.11.

1.2.9 Other Data Visualisation Methods

The above sections have presented the most common data visualisation methods for a given data set. More complex forms can be created by combining different simple data visualisation methods into a final integrated plot. Alternatively, the data could be transformed (changed in some manner) before being plotted. The different techniques that are available to accomplish this depend strongly on the intended application and will be introduced in the relevant sections in later chapters. Often such plots are created when there is multiple information that needs to be displayed,

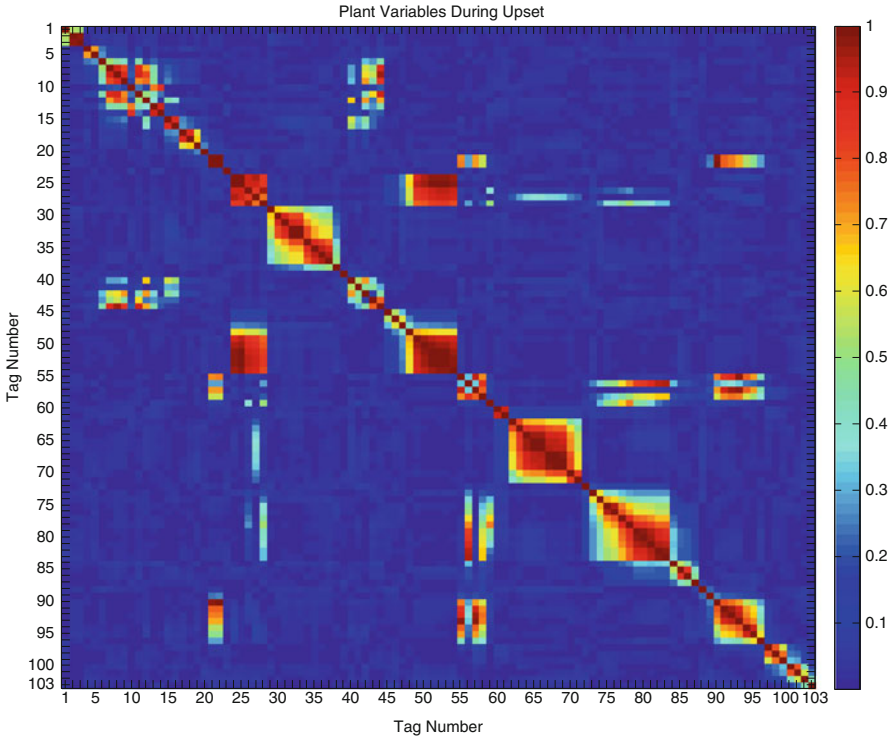


Fig. 1.12 Complex data visualisation example: a cross-correlation plot

for example, one is interested in determining which of 20 variables are important for your analysis.

Two typical integrated data visualisation methods are presented in Figs. 1.12 and 1.13. In Fig. 1.12, the linear relationship between 100 different variables is plotted to determine which variables are most related with each other. This plot involves taking the data, transforming it, and then computing the correlation between each pair of the transformed data. A strong linear relationship is denoted by 1 (or a red colour), while a weak linear correlation is denoted by 0 (or a dark blue colour). Obviously, the variables themselves are strongly related with each other and so the diagonal is always equal to 1 in such plots. More information on creating and plotting such figures can be found in Chap. 5. In Fig. 1.13, two variables are plotted against each other as a scatter plot with histograms to show the distribution of the individual variables. These plots can be useful for seeing and understanding complex interactions between different variables and how best to interpret them later. In this particular example, it can be seen that both variables are skewed to the left, with a rather large right tail.

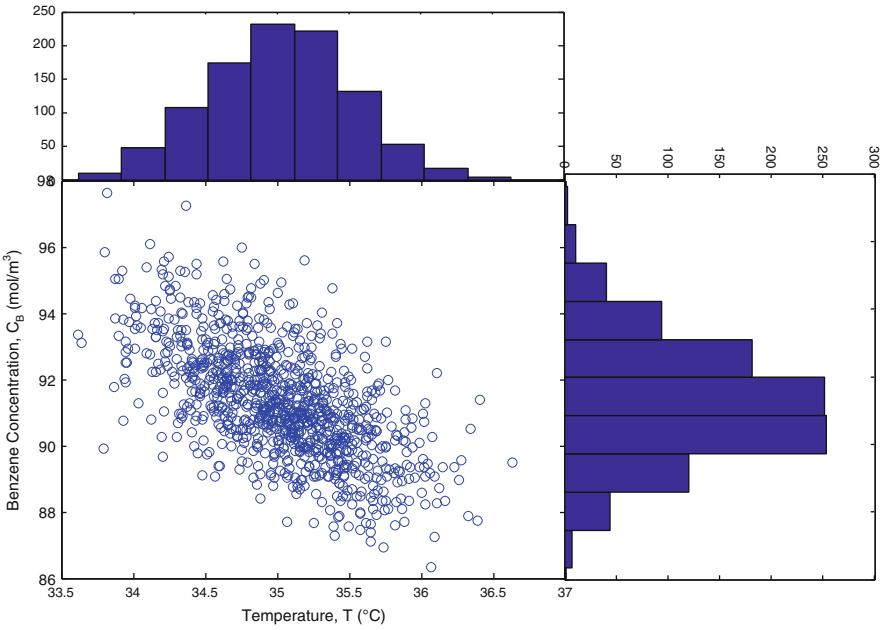


Fig. 1.13 Complex data visualisation example: combining multiple plot types

1.3 Friction Factor Example

In this section, experimental data from a friction factor experiment will be considered. This data set consists of four separate runs performed on different pipe diameters collected on different days (often with a large separation in time).

1.3.1 Explanation of the Data Set

In the friction factor experiment, the flow of water through a pipe is changed to determine the pressure drop across a length of pipe for pipes with different diameters. In order to compare the results across multiple different diameters, the data are converted into two dimensionless numbers: the Reynolds number (Re), which represents the flow and is defined as

$$\mathrm{Re} = \frac{\rho v D}{\mu} \tag{1.14}$$

Table 1.4 Data from friction factor experiments

Run 1		Run 2		Run 3		Run 4	
Re	f	Re	f	Re	f	Re	f
6,478	0.0355	19,476	0.0268	20,701	0.0251	11,529	0.0308
11,785	0.0303	13,439	0.0293	13,248	0.0286	9,993	0.0318
5,485	0.0369	15,844	0.0281	18,409	0.0266	9,340	0.0329
9,075	0.0321	5,251	0.0369	5,602	0.0351	3,187	0.0420
11,815	0.0302	11,980	0.0303	14,251	0.0281	6,248	0.0362
7,246	0.0343	17,732	0.0272	18,978	0.0261	4,838	0.0387
10,403	0.0309	6,366	0.0352	9,787	0.0309	4,427	0.0394
13,364	0.0292	15,115	0.0283	6,638	0.0339	9,567	0.0327
10,811	0.0310	7,461	0.0345	10,748	0.0302	7,141	0.0351
7,730	0.0334	10,227	0.0314	16,813	0.0270	5,750	0.0371
9,938	0.0316	13,240	0.0296	12,730	0.0290	11,187	0.0312
11,581	0.0305	13,987	0.0291	8,794	0.0319	3,925	0.0405
8,432	0.0327	16,606	0.0277	15,041	0.0278		
12,546	0.0297	11,152	0.0307	12,060	0.0292		
9,051	0.0325	5,226	0.0377	6,937	0.0337		
9,470	0.0317			4,895	0.0364		

where ρ is the density of the fluid, v is the velocity, D is the pipe diameter, and μ is the dynamic viscosity of the fluid, and the friction factor (f), which represents the pressure drop in the pipe and is defined as

$$f = \frac{2D\Delta P}{\rho v^2 L} \tag{1.15}$$

where L is the length of the pipe and ΔP is the pressure drop.

The relationship between the friction factor and Reynolds number can be written as (Gerhart et al. 1992):

$$f = K\mathrm{Re}^{\beta} \tag{1.16}$$

where K and β are parameters to be fit. For turbulent flow, where $4{,}000 < \mathrm{Re} < 100{,}000$, the Blasius equation predicts that $K = 0.316$ and $\beta = -0.25$ (Gerhart et al. 1992).

The experiment consisted of data collected on multiple days for different pipe diameters and flow rates using water as the fluid. Sample data are presented in Table 1.4. Runs 1 and 2 were performed on the same day, but with different pipe diameters: 4.9 mm for Run 1 and 6.1 mm for Run 2. Run 3 was performed on another day with a pipe diameter of 7.8 mm. Finally, Run 4 was some historical data obtained 6 years previously using the same equipment and a pipe diameter of 4.9 mm. The data are presented sequentially in the order in which the experiments were run, that is, for example, in Run 1, the experiment with a $\mathrm{Re} = 6{,}478$ was run

Table 1.5 Summary statistics for the friction factor data set

	Run 1		Run 2		Run 3		Run 4	
Summary statistic	Re	f	Re	f	Re	f	Re	f
Mean	9,700	0.0320	12,200	0.0309	12,200	0.0300	7,260	0.0357
Median	9,700	0.0317	13,200	0.0296	12,400	0.0291	6,700	0.0357
σ	2,300	0.0021	4,500	0.0036	4,900	0.0034	2,900	0.0039
Range	7,880	0.0077	14,300	0.0109	15,800	0.0113	8,340	0.0112
σ_{MAD}	1,900	0.0012	3,000	0.0018	4,000	0.0023	2,700	0.0034

first, followed by the experiment with Re = 117,785. Replicates were performed at some of the values, for example, in Run 1, there are two cases with a Re ≈ 11,800.

1.3.2 Summary Statistics

The mean, median, standard deviation, range, and median absolute difference will be determined for all four runs. Sample computations will be shown for Run 4 using the Reynolds number values. The results are summarised in Table 1.5.

For Run 4 and the Re values, the mean would be computed using Eq. (1.1) to give

$$\begin{aligned}\overline{\text{Re}} &= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\begin{array}{l}11{,}529 + 9{,}993 + 9{,}340 + 3{,}187 + 6{,}248 + 4{,}838 \\ + 4{,}427 + 9{,}567 + 7{,}141 + 5{,}750 + 11{,}187 + 3{,}925\end{array}}{12} \\ &= 7{,}261\end{aligned} \tag{1.17}$$

Similarly, the median would be computed by first ordering the data set from smallest to largest and then finding the average of the two midpoint values (since there is an even number of values present), that is,

$$3{,}187; 3{,}925; 4{,}427; 4{,}838; 5{,}750; \underbrace{6{,}248; 7{,}141}; 9{,}340; 9{,}567; 9{,}993; 11{,}187; 11{,}529$$

$$\boxed{\begin{aligned}\text{median} &= \frac{6{,}248 + 7{,}141}{2} \\ &= 6{,}694.5\end{aligned}} \tag{1.18}$$

The standard deviation can be computed using a modified form of Eq. (1.3) commonly used for manual computations to give

$$\begin{aligned}\sigma_{\mathrm{Re}} &= \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}} = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}} \\ &= \sqrt{\frac{\left(\begin{array}{l} 11{,}529^2 + 9{,}993^2 + 9{,}340^2 + 3{,}187^2 + 6{,}248^2 + 4{,}838^2 \\ + 4{,}427^2 + 9{,}567^2 + 7{,}141^2 + 5{,}750^2 + 11{,}187^2 + 3{,}925^2 \end{array}\right) - 12(7{,}261)^2}{12-1}} \\ &= 2{,}900\end{aligned} \tag{1.19}$$

The range can be found by determining the largest and smallest values and subtracting them. Thus, the maximum value is 11,529 and the minimum value is 3,187. Therefore, the range is $11{,}529 - 3{,}187 = 8{,}340$.

The median absolute difference can be computed by first ordering the absolute value of the difference between the data point and the median to give

$$3{,}507.5; 2{,}769.5; 2{,}267.5; 1{,}856.5; 944.5; 446.5; 446.5; 2{,}645.5; 2{,}872.5; 3{,}298.5; 4{,}492.5; 4{,}834.5 \tag{1.20}$$

The ordered list then becomes

$$446.5; 446.5; 944.5; 1{,}856.5; 2{,}267.5; \underbrace{2{,}645.5; 2{,}769.5}_{\boxed{\text{median} = 2{,}707.5}}; 2{,}872.5; 3{,}298.5; 3{,}507.5; 4{,}492.5; 4{,}834.5 \tag{1.21}$$

The median of the residuals is therefore 2,707.5.

It should be noted that all of the values have been rounded to three decimal places, except for the standard deviation, which has been rounded to two decimal places, in order to improve the presentation. It should be noted that the original mass flow rates and pressure drops used to compute the Reynolds number and friction factor were recorded to only three decimal places.

1.3.3 Data Visualisation

In this particular case, a scatter plot showing all the 4 runs together and a box-and-whisker plot of each run separately will be plotted. Detailed code for creating these graphs is given in Chap. 7 for MATLAB® and Chap. 8 for Microsoft Excel®. Figure 1.14 shows a scatter plot of the data showing each of the runs separately, while Fig. 1.15 gives the box-and-whisker plots for both the Reynolds number and the friction factor. The theoretical values using the Blasius equation have also been included in Fig. 1.14 to provide some reference point against which to compare the data set.

In order to illustrate the procedure for constructing a box-and-whisker plot by hand and determining the appropriate quartile boundaries, the Reynolds numbers

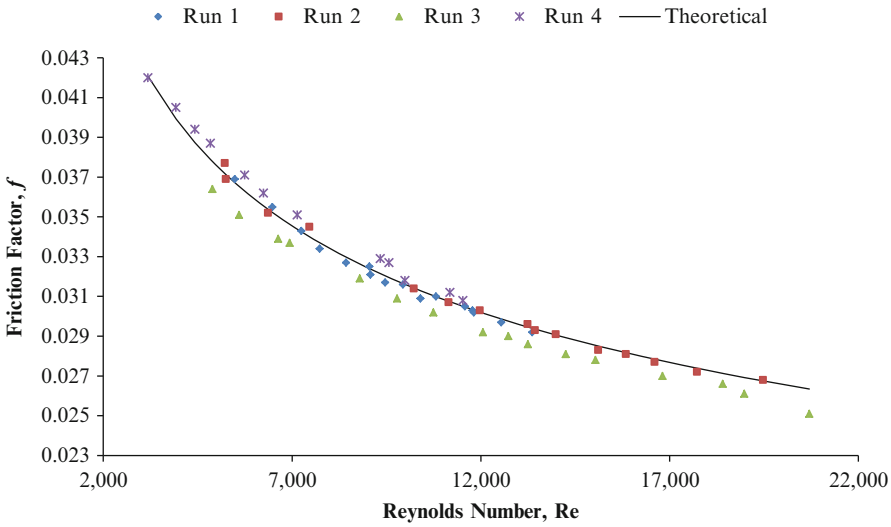


Fig. 1.14 Scatter plot of the friction factor as a function of Reynolds number for all four runs

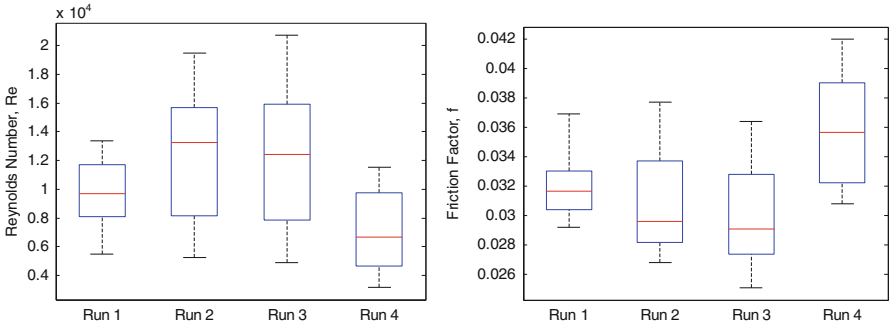


Fig. 1.15 Box-and-whisker plots for the friction factor experiment for the (*left*) Reynolds number and (*right*) friction factor

from Run 4 will be used. For a box-and-whisker plot, it is necessary to determine the values located at $Q_{0.25}$, $Q_{0.5}$ (= median) and $Q_{0.75}$. Equation (1.7) gives a general formula for computing these values. For $Q_{0.25}$, the first quartile, setting $n = 12$ and $p = 0.25 = ¼$ in the formula gives:

$$\begin{aligned} h &= (12-1)0.25 + 1 = \frac{15}{4} = 3.75 \\ Q_{0.25} &= x_{\lfloor 3.75 \rfloor} + (3.75 - \lfloor 3.75 \rfloor)\left(x_{\lfloor 3.75 \rfloor + 1} - x_{\lfloor 3.75 \rfloor}\right) \\ &= x_3 + (3.75 - 3)(x_4 - x_3) \\ &= 4{,}427 + 0.75(4{,}838 - 4{,}427) \\ &= 4{,}735 \end{aligned} \tag{1.22}$$

Similarly, $Q_{0.5}$ can be computed as

$$\begin{aligned} h &= (12-1)0.5+1 = \frac{13}{2} = 6.5 \\ Q_{0.5} &= x_{\lfloor 6.5 \rfloor} + (6.5 - \lfloor 6.5 \rfloor)\left(x_{\lfloor 6.5 \rfloor + 1} - x_{\lfloor 6.5 \rfloor}\right) \\ &= x_6 + (0.5)(x_7 - x_6) \\ &= 6{,}248 + 0.5(7{,}141 - 6{,}248) \\ &= 6{,}695 \end{aligned} \tag{1.23}$$

It can be noted that, after rounding, this value is identical to that previously computed for the median. This should be always the case.

Similarly, $Q_{0.75}$ can be computed as

$$\begin{aligned} h &= (12-1)0.75+1 = \frac{37}{4} = 9.25 \\ Q_{0.75} &= x_{\lfloor 9.25 \rfloor} + (9.25 - \lfloor 9.25 \rfloor)\left(x_{\lfloor 9.25 \rfloor + 1} - x_{\lfloor 9.25 \rfloor}\right) \\ &= x_9 + (0.25)(x_{10} - x_9) \\ &= 9{,}567 + 0.25(9{,}993 - 9{,}567) \\ &= 9{,}674 \end{aligned} \tag{1.24}$$

For comparison, the values computed above are compared with the values obtained using different software in Table 1.6. It can be seen that each software package can compute the same value differently. In all cases, the median will be computed the same way, since it is a fixed value. As was previously mentioned, this verifies that `quartile.inc` function in Excel is equivalent to the values obtained manually based on Eq. (1.7), while `quartile.exc` function in Excel is based on option 1 for Eq. (1.7). Finally, MATLAB uses option 2 for Eq. (1.7). Nevertheless, all values are relatively close to each other and would not impact too greatly the overall results.

1.3.4 Some Observations on the Data Set

First, consider the results in Table 1.5, which presents the summary statistics for the data set. It can be noted that for Runs 2 and 3, which both have a similar mean Reynolds Number, the median is quite different for each. This suggests that the distribution is different. Looking at Fig. 1.15 for these two runs, it can be seen that Run 3 has more extreme values (in both directions) than Run 2, which will balance out both the mean and median values. On the other hand, Fig. 1.15 shows that for Run 2, the size of the Q2–Q3 area is much smaller than for Run 3, suggesting that 25% of the data are compactly located in a small area. On the other hand, for Run 1, the mean and median are more closely aligned, which suggests that the data are more evenly distributed. This is confirmed by looking at Fig. 1.15 for Run 1, where the size of the two boxes is almost equal. Run 4 for the friction factor has a similar even distribution. In all cases, Table 1.5 shows that a larger range implies that the standard deviation will also be larger.

Table 1.6 Computing quartiles with different software packages

Quartile	Manual	Excel® 2010 (`quartile.inc`)	Excel® 2010 (`quartile.exc`)	MATLAB® 2014
1	4,735	4,735	4,530	4,633
2	6,695	6,695	6,695	6,695
3	9,674	9,674	9,887	9,780

Next, consider the scatter plot shown in Fig. 1.14, where a scatter plot of the data by run and the theoretical values are presented. Note that each run is denoted by a symbol that appears distinct even if there is no colour. From here, it can be observed that Run 3 is consistently below the theoretical value. This suggests that this run could potentially be some sort of outlier. Furthermore, Run 4 seems to have been performed at much lower Reynolds numbers than the rest of the experiments. This difference is even evident from the summary statistics.

1.4 Further Reading

The following are references that provide additional information about the topic:

1. *History of Statistics:*
   (a) Hald A (2003) A history of probability and statistics and their application before 1750. Wiley, Hoboken
   (b) Sheynin O (2004) History of the theory of probability to the beginning of the 20th century. NG Verlag, Berlin
   (c) Varberg DE (1963) The development of modern statistics. Math Teach 56 (4):252–257
2. *Data Analysis*:
   (a) Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, Chichester
   (b) Daniel C, Wood FS (1980) Fitting equations to data, 2nd edn. Wiley, New York
   (c) Davies L, Gather U (1993) The identification of multiple outliers. J Am Stat Assoc 88(423):782–792
   (d) Hawkins DM (1980) Identification of outliers. Chapman and Hall, London
   (e) Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22:85–126
   (f) Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. Am Stat 50(4):361–365
   (g) Lin B, Recke B, Knudsen JK, Jørgensen SB (2007) A systematic approach for soft sensor development. Comput Chem Eng 31:419–425

3. *Data Visualisation*:

(a) Tufte ER (1997) Visual and statistical thinking: displays of evidence for making decisions. Graphics Press LLC, Cheshire
(b) Tufte ER (2001) The visual display of quantitative information. Graphics Press LLC, Cheshire

1.5 Chapter Problems

Problems at the end of the chapter consist of three different types: (a) Basic Concepts (True/False), which seek to test the reader's comprehension of the key concepts in the chapter; (b) Short Exercises, which seek to test the reader's ability to compute the required parameters for a simple data set using simple or no technological aids, and this section also includes proofs of theorems; and (c) Computational Exercises, which require not only a solid comprehension of the basic material but also the use of appropriate software to easily manipulate the given data sets.

1.5.1 Basic Concepts

Determine if the following statements are true or false and state why this is the case.

1. The mean is a robust measure of central tendency.
2. A trimodal data set has four modes.
3. The median measures the middle value of a data set.
4. The median and the mean will always be the same.
5. The variance is equal to the standard deviation squared.
6. The range is a useful measure of the spread of the data.
7. The median absolute difference is a robust measure of dispersion.
8. A left-skewed data set has many values in the left tail.
9. The skewness of a data set measures how symmetric the data set is.
10. Sextiles partition a data set into six parts.
11. Outliers are data points whose values are abnormal.
12. A graph should have clearly labelled axes and an appropriate legend.
13. Graphs containing many different symbols distinguished solely by colour are well designed.
14. Pie charts are the foundation upon which histograms are constructed.
15. Sparkplots are useful for describing trends and general behaviour of a data set.
16. Tables are useful for summarising important information, such as mean and variance, of a data set.
17. Taking a numeric value directly from software and placing it unformatted into a table is a good idea.

18. A probability plot is useful for comparing the data set against some theoretical distribution.
19. Transforming a data set can lead to a more meaningful graph.
20. Combining different types of graphs together can create a graph with more information.

1.5.2 Short Exercises

These questions should be solved using only a simple, nonprogrammable, nongraphical calculator combined with pen and paper.

21. For the data set $\{1, 3, 5, 2, 5, 7, 5, 2, 8, 5\}$,

    (a) Compute the mean, mode, and median.
    (b) Compute the variance, median absolute difference, and range.
    (c) Compute the first, second, and third quartiles.
    (d) Plot a box-and-whisker plot.
    (e) Plot a histogram with bins $x < 2$, $2 \leq x < 4$, $4 \leq x < 6$, $6 \leq x < 8$, and $x \geq 8$.

22. For the data set $\{2.3, 1.2, 3.4, 4.5, 3.4, 1.2, 3.4, 4.0, 1.1\}$,

    (a) Compute the mean, mode, and median.
    (b) Compute the variance, median absolute difference, and range.
    (c) Compute the first, second, third, and fourth quintiles.
    (d) Plot a box-and-whisker plot.
    (e) Plot a histogram with bins $x < 2$, $2 \leq x < 3$, $3 \leq x < 4$, and $x \geq 4$.

1.5.3 Computational Exercises

The following problems should be solved with the help of a computer and appropriate software packages, such as MATLAB® or Excel®.

23. Consider the data in Table 1.7 that shows the different faults (problems) associated with running a reactor over a 30-day period. A fault can occur multiple times in a given time frame. Compute appropriate summary statistics and create appropriate graphs to summarise the data. (*Hint: there is no one single correct solution.*)
24. Consider the data in Table 1.8 that shows the flow rate of steam in kg/h through a pipe. Due to the presence of stiction and other nonlinearities in the control valve, a new control algorithm is being proposed. The engineer in charge of making the change has to evaluate whether the new algorithm is better. A better algorithm is defined as one that reduces the variance of the steam flow rate and can keep the process closer to the desired set point of 8.5 kg/h. The original and new control methods are both tested for 2 h and the data are collected every

Table 1.7 Reactor fault types by shift (for Question 23)

Fault type	Number of faults by shift			
	Night (midnight to 6:00 a.m.)	Morning (6:00 a.m. to noon)	Afternoon (noon to 6:00 p.m.)	Evening (6:00 p.m. to midnight)
High reactor level	5	6	2	6
Abnormal pressure	10	2	2	5
Explosion	2	0	0	0
Low temperature	5	2	10	5
High temperature	5	8	0	10
Others	2	10	5	0

Table 1.8 Steam control data with two different methods (for Question 24)

Time (min)		5	10	15	20	25	30	35	40	45	50	55	60
Base	1 h	8.5	8.7	8.4	8.6	8.2	8.7	8.9	8.5	8.5	8.4	8.3	8.6
	2 h	8.2	8.4	8.3	8.2	8.4	8.5	8.8	8.3	8.6	8.7	8.5	8.3
New	1 h	8.4	8.5	8.4	8.5	8.6	8.3	8.6	8.7	8.2	8.3	8.4	8.5
	2 h	8.5	8.6	8.4	8.3	8.4	8.6	8.7	8.5	8.5	8.5	8.3	8.4

5 min. Plot the available data and analyse it. Without using any formal statistical tests, suggest whether the proposed control algorithm is better than the original, base case.

25. Take any large data set that is of interest to you and analyse it using the methods presented in this chapter. The data set should have at least 1,000 data points and two variables. You can then use this data set in subsequent chapters to perform additional analysis.